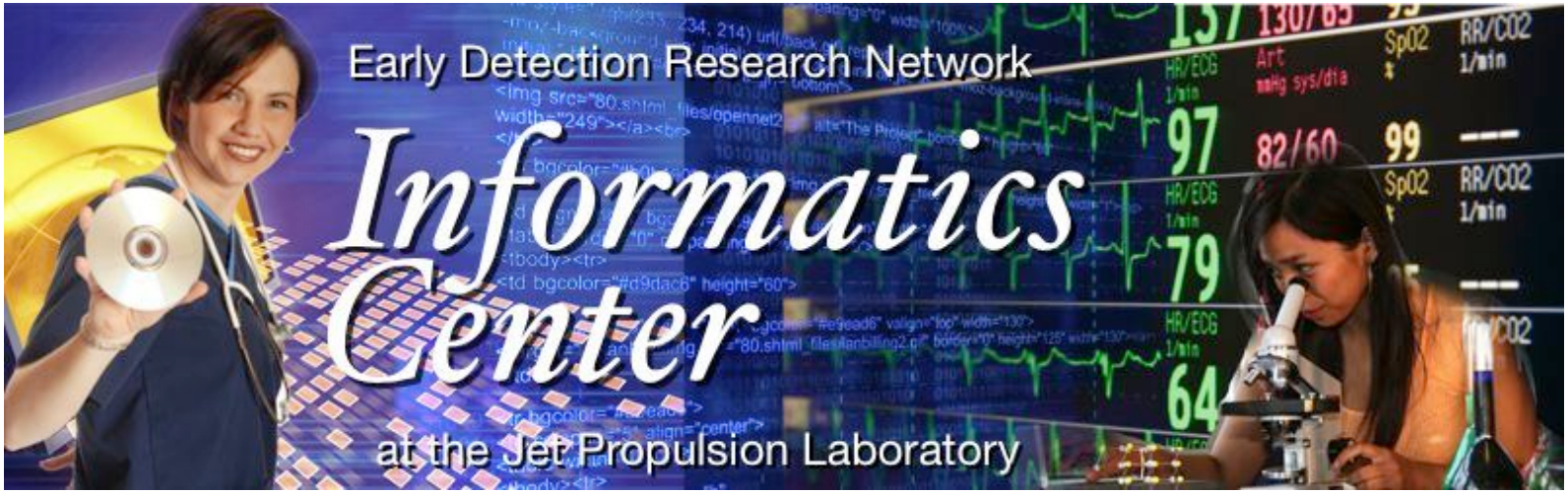


Overview of the EDRN Informatics Center: eDRN Catalog and Archive Service

Chris A. Mattmann

Schedule

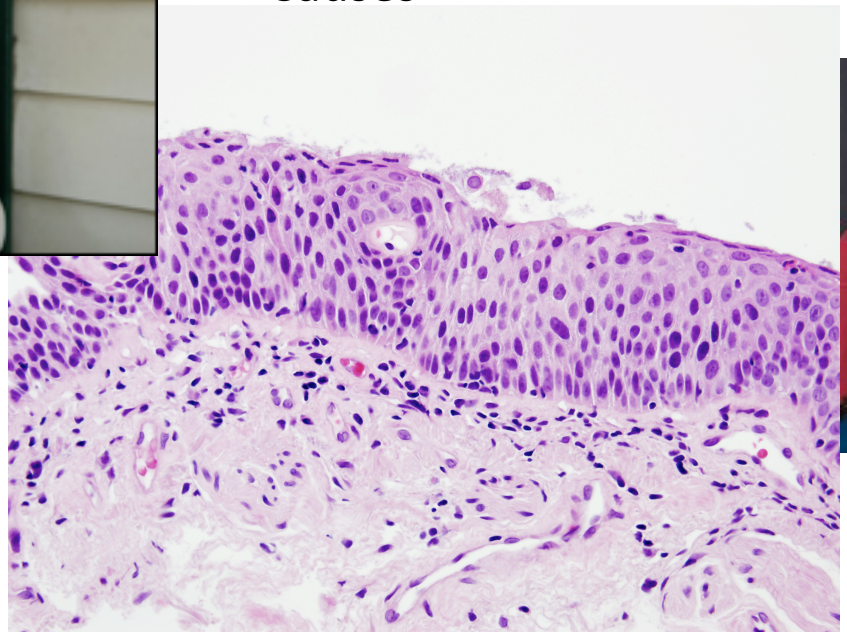
- What is the EDRN IC?
- Who's involved?
- What are the types of applications we build?
- How do we build them?
- Focus: eCAS
- Results to date
- Where we're headed

The Early Detection Research Network

- EDRN is a network of 40+ institutions all performing research geared towards the discovery of cancer *biomarkers*, which are early indicators of onset of disease
- NCI/NIH funded program
  - 100s of millions of dollars of investment
  - Recently renewed after NIH Board of Scientific Advisors advised NIH of the strategic value in EDRN and the existing successes

Cancer Biomarkers

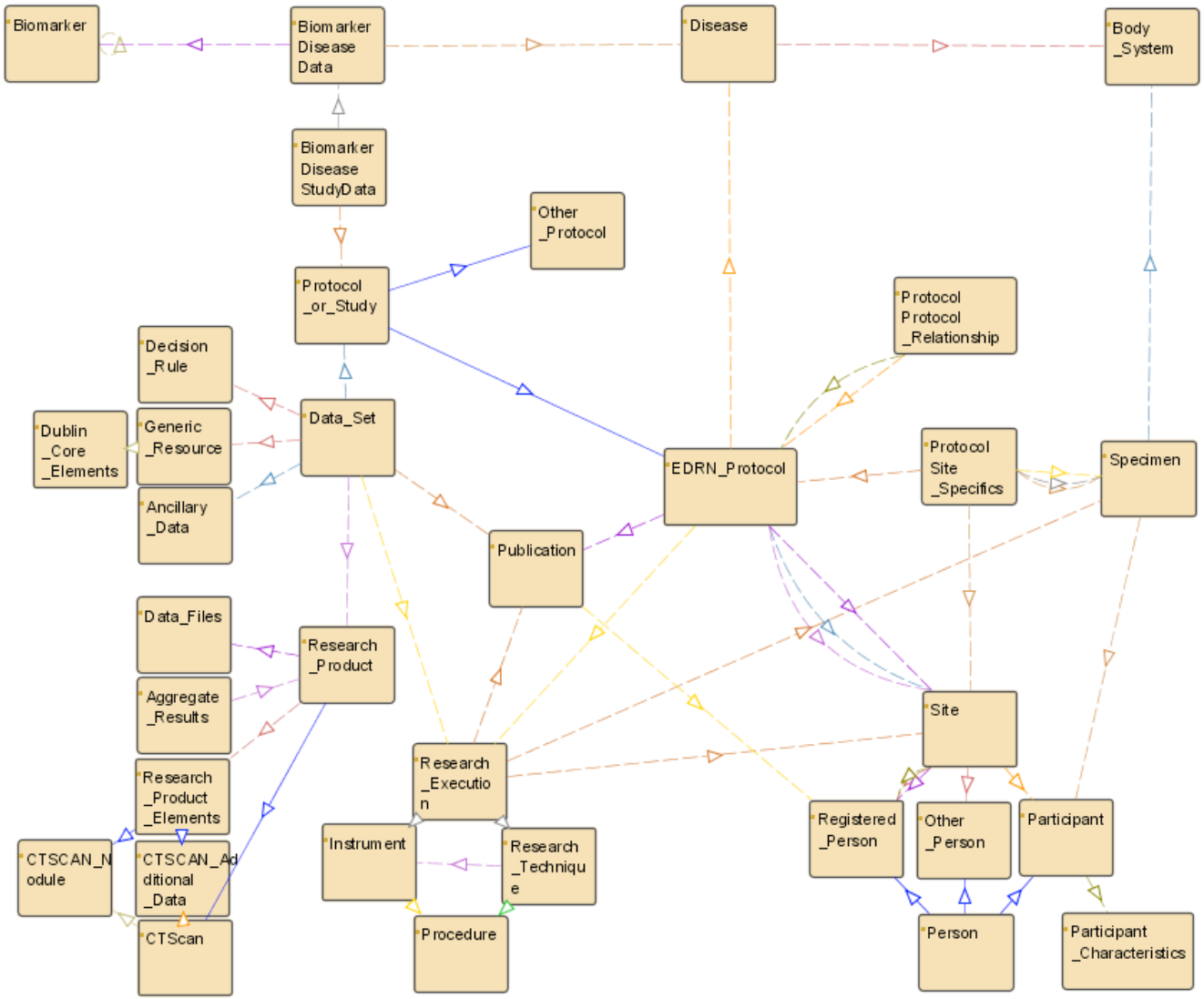
- Indicators of early onset of disease
  - Important to capture information about:
    - Characteristics of patients and onsets (epidemiological information)
    - Causes



Types of information important in EDRN

- Specimen Inventories
  - At each site, maps specimens collected (blood, sputum, etc.) to patient characteristics
- Studies and Protocols
  - Information about studies conducted in the EDRN and results (publications, outputs)
- Biomarkers
  - Information about indicators of early disease
- Science Data
  - Outputs of experiments on specimens, regarding biomarkers, driven by particular studies and protocols

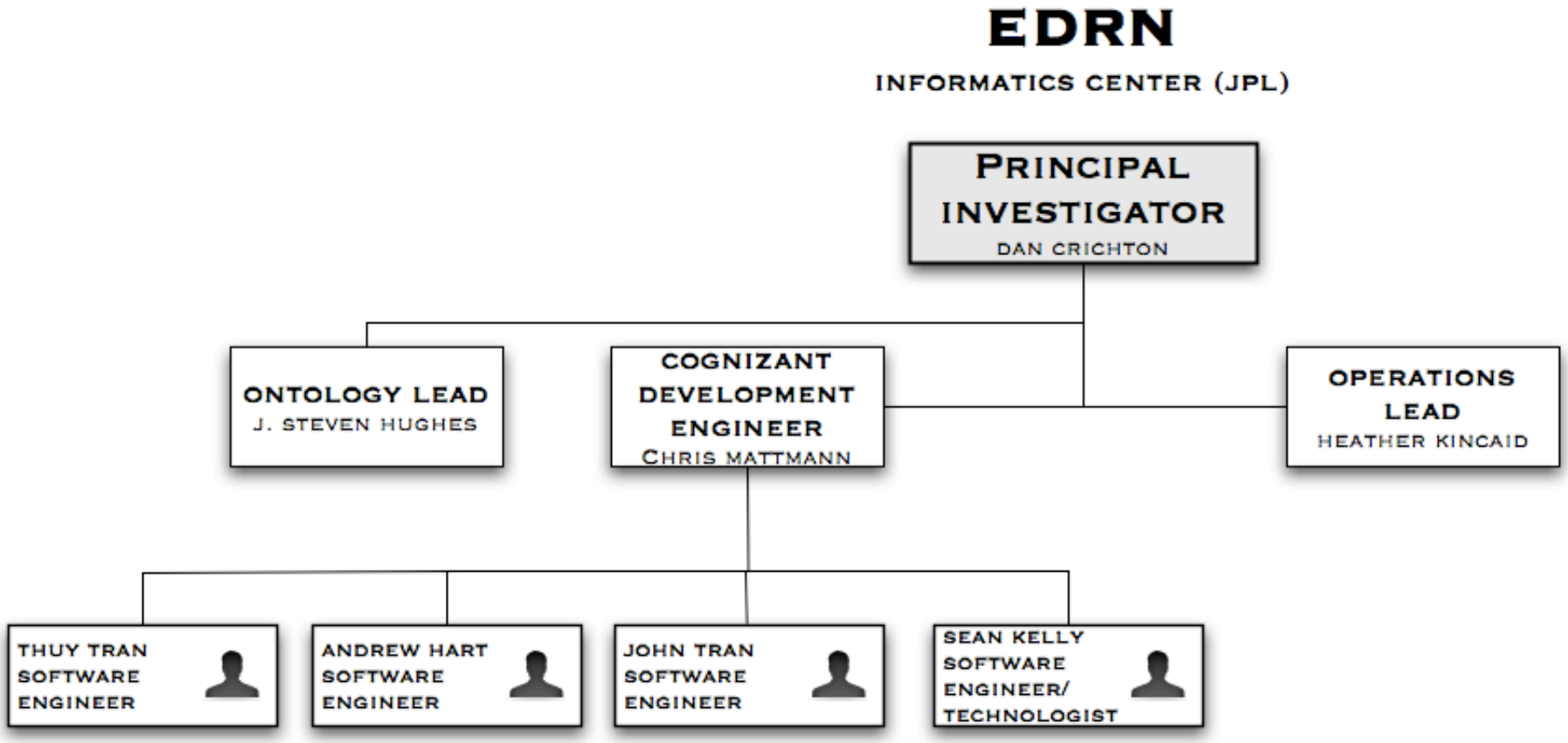
Types of information important in EDRN



The EDRN Informatics Center's Role

- The EDRN Informatics Center develops software solutions to support the Early Detection Research Network's research of cancer biomarkers and development of cancer-fighting tools.
  - Develop software solutions and provide informatics expertise to assist EDRN centers and NCI in capturing and disseminating important EDRN information mentioned before

Informatics Center Organization



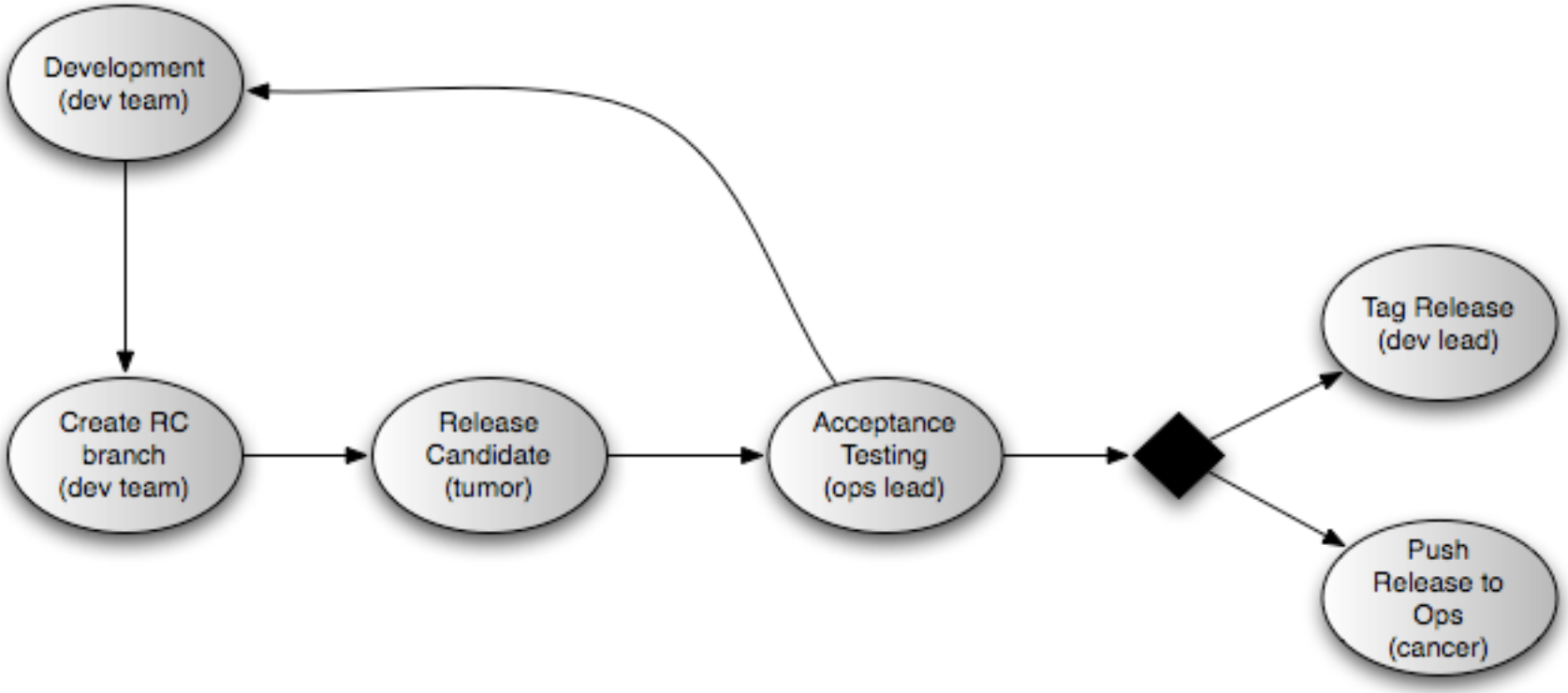
EDRN Informatics Center Hardware

- Tumor
  - FusionX5 1U High-performance Node w/ 1x 5140 Dual-core CPUs
  - 0.5 TB total space
  - Development machine

- Cancer
  - FusionX5 1U High-performance Node w/ 2x 5140 Dual-core CPUs
  - 1 TB total space
  - Operations machine

EDRN Informatics Center: Development Process

- *Read:* agile



Use of modern software process tools

Issue tracking

Source code repository

EDRN IC site

Repository browsing

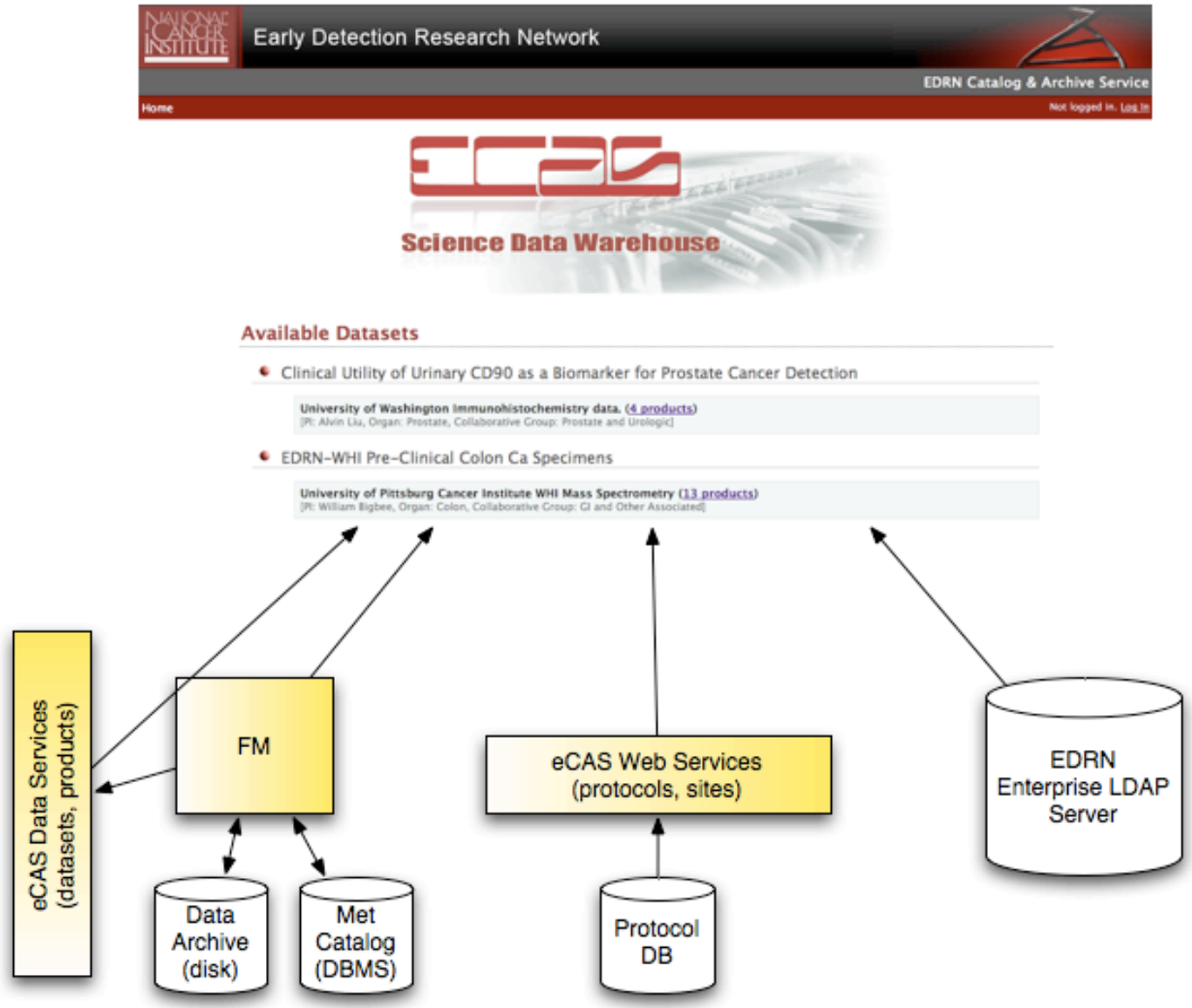
Mailing lists

Acceptance testing

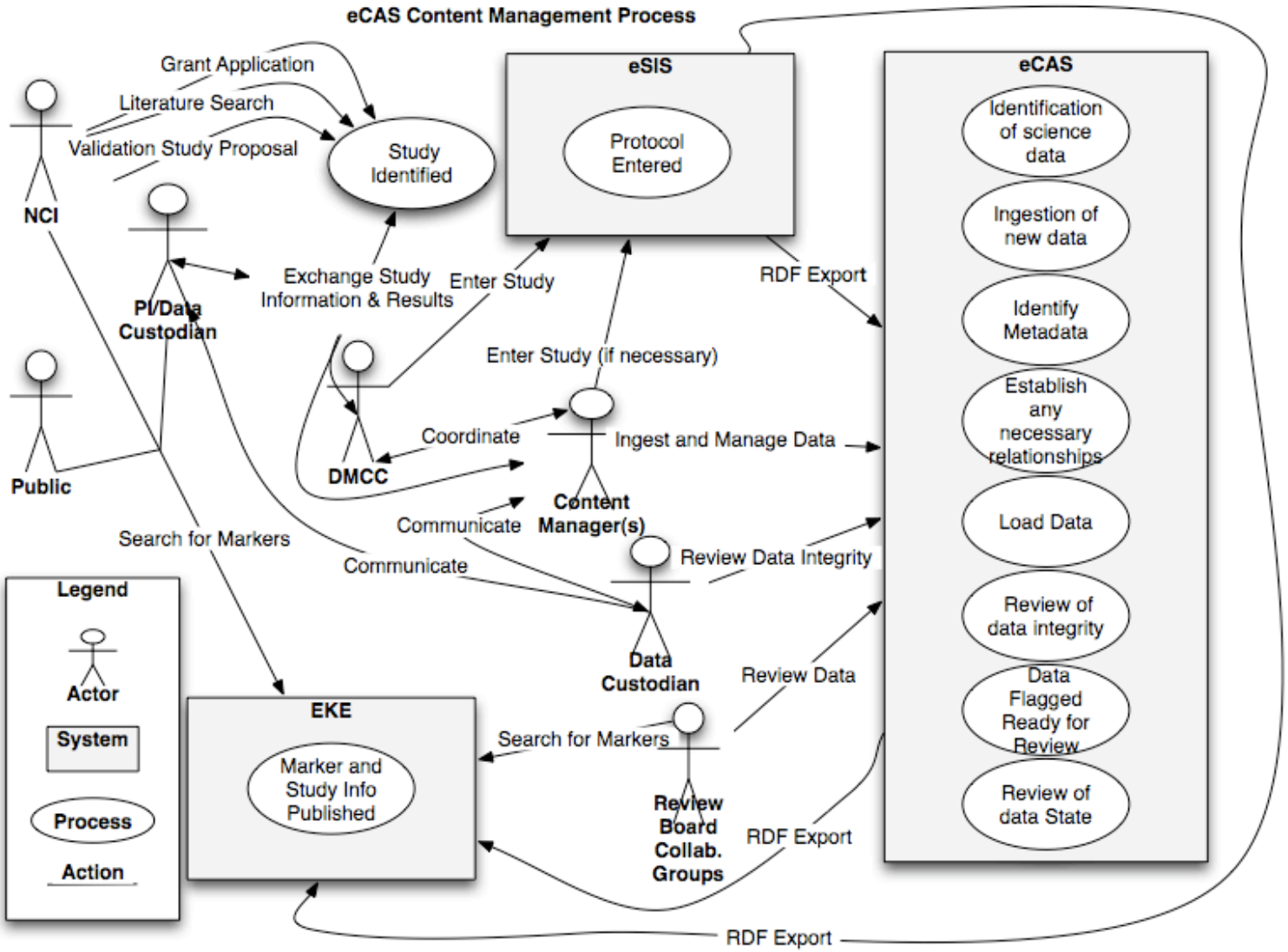
Let's talk about eCAS

- Data Warehousing Application
  - Store raw and processed data from local laboratories
  - NCI eCAS that holds curated "high quality" eCAS data published from local sites
- Many potential applications
  - Specimen images, datasets, instrument generated products, etc.

eCAS High Level Architecture



eCAS Curation Process

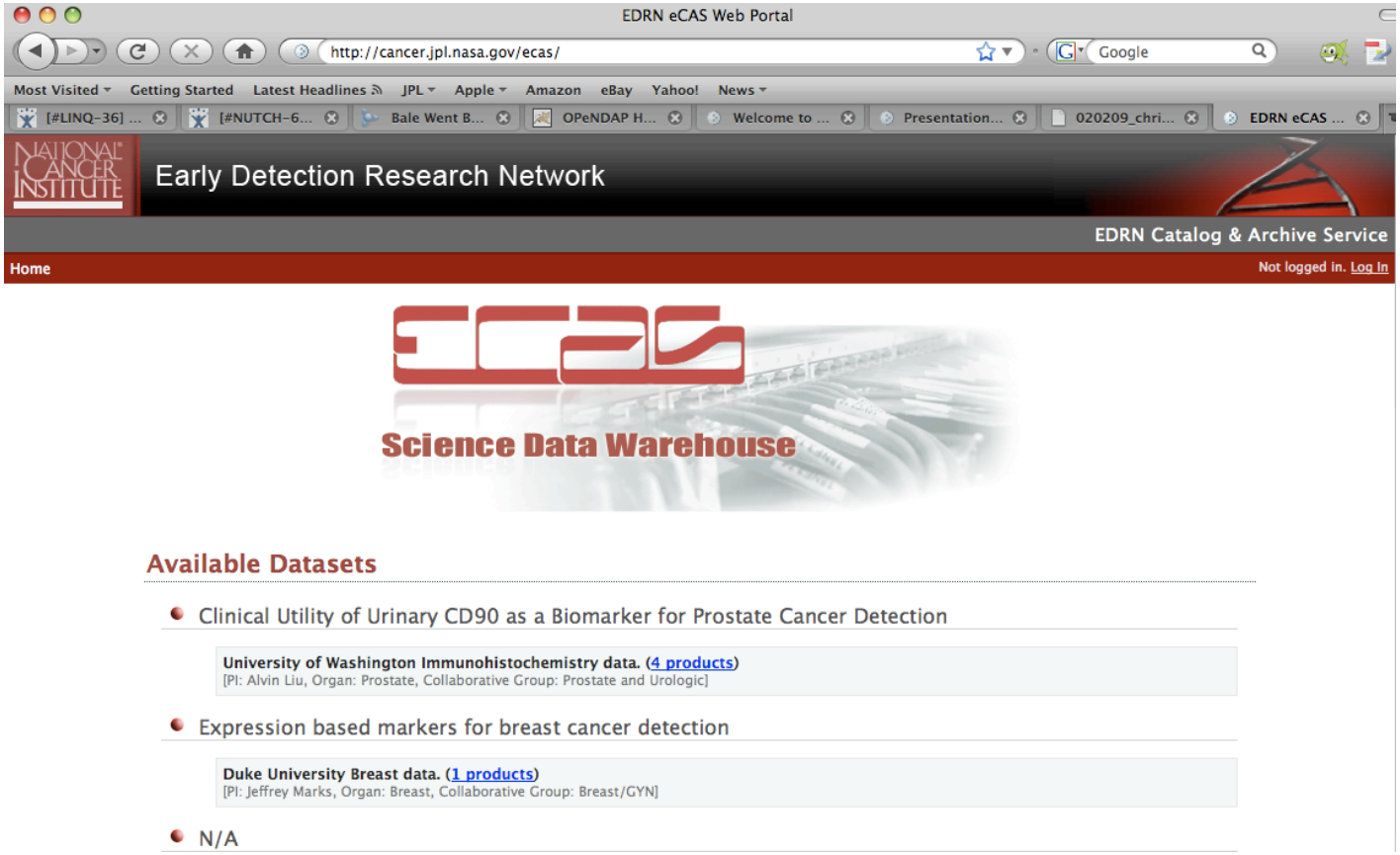


Role of DMCC and NCI

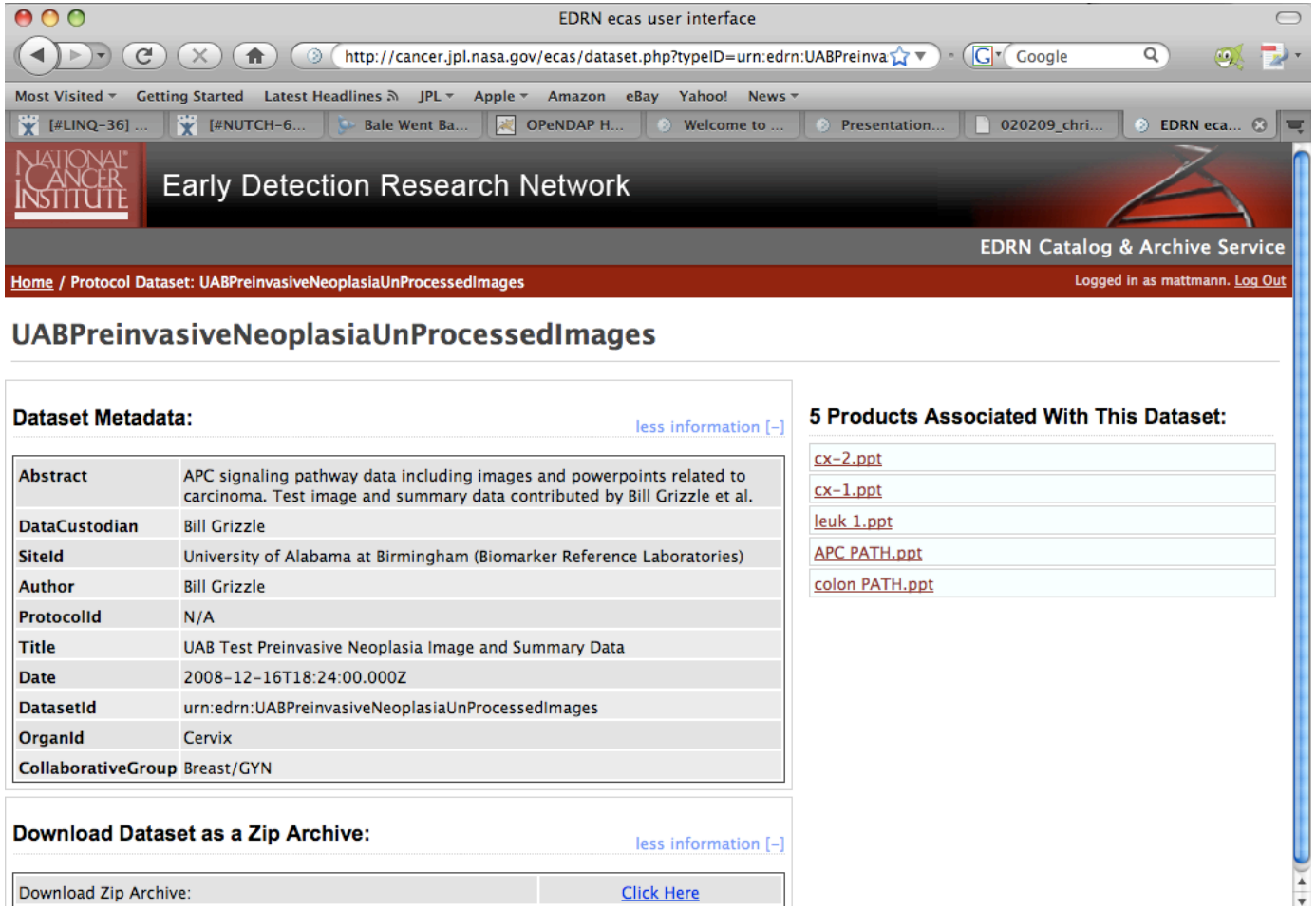
- DMCC
  - Help to derive metadata standards for different dataset types
  - Annotation of datasets
  - Delivery of data (processed)
- NCI
  - Encourage PIs to submit data to eCAS
  - Encourage PIs to run local laboratory eCAS
  - Capture Prostate 2000 data along with disclaimer

eCAS Interface: Splash Page

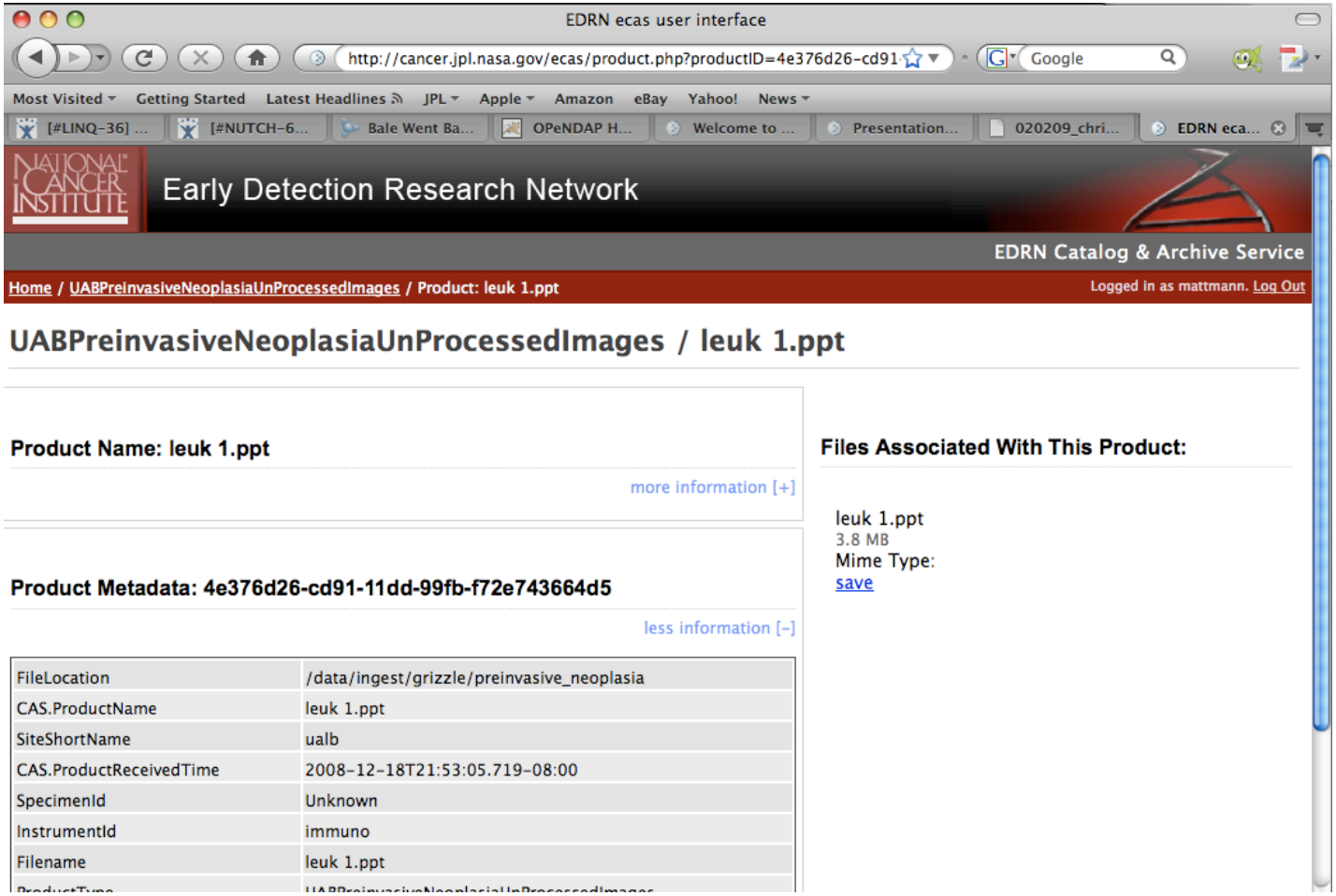
- http://ecas.jpl.nasa.gov
  - Operational 0.4.2 beta release



eCAS Interface: Logged in Dataset View



eCAS Interface: Logged in Product View



Capabilities

- Current (as of 0.4.2 beta)
  - Download datasets as zip files
    - Zips include zips of products (and their metadata), as well as dataset-level metadata
  - Download products as zip files
  - Browse of dataset and product metadata
  - Dump RDF dataset and product information into EKE
  - Pagination of results
- Planned (in next 2 months, 2-4 releases)
  - Curation interface
    - Talk more about this later, with John's slides
  - Bug fixes/maintenance

Current Datasets

- SELDI Phase I and II,MALDI Dilution Data, Miscellaneous PI data, (Immunohistochemistry, LabMAP Mass Spectrometry, FISH MSA data), WHI datasets
- Test data from Bill Grizzle (as of 0.3.0 beta release)
  - EGFR Translocation Data
  - Pre-invasive Neoplasia Image and Summary Data
- Sites
  - NIST, FHCRC, EVMS, UAB, Duke, Univ Pitt, Univ Wash, Wayne State University

Security

- Current
  - Integration with EDRN Informatics Center LDAP server
  - Basic authentication, to guard datasets and products, and download
- Planned
  - Granular authorization, based on roles
    - Need to develop dataset, product-level policy, as well as metadata attribute browse, and downloading policy

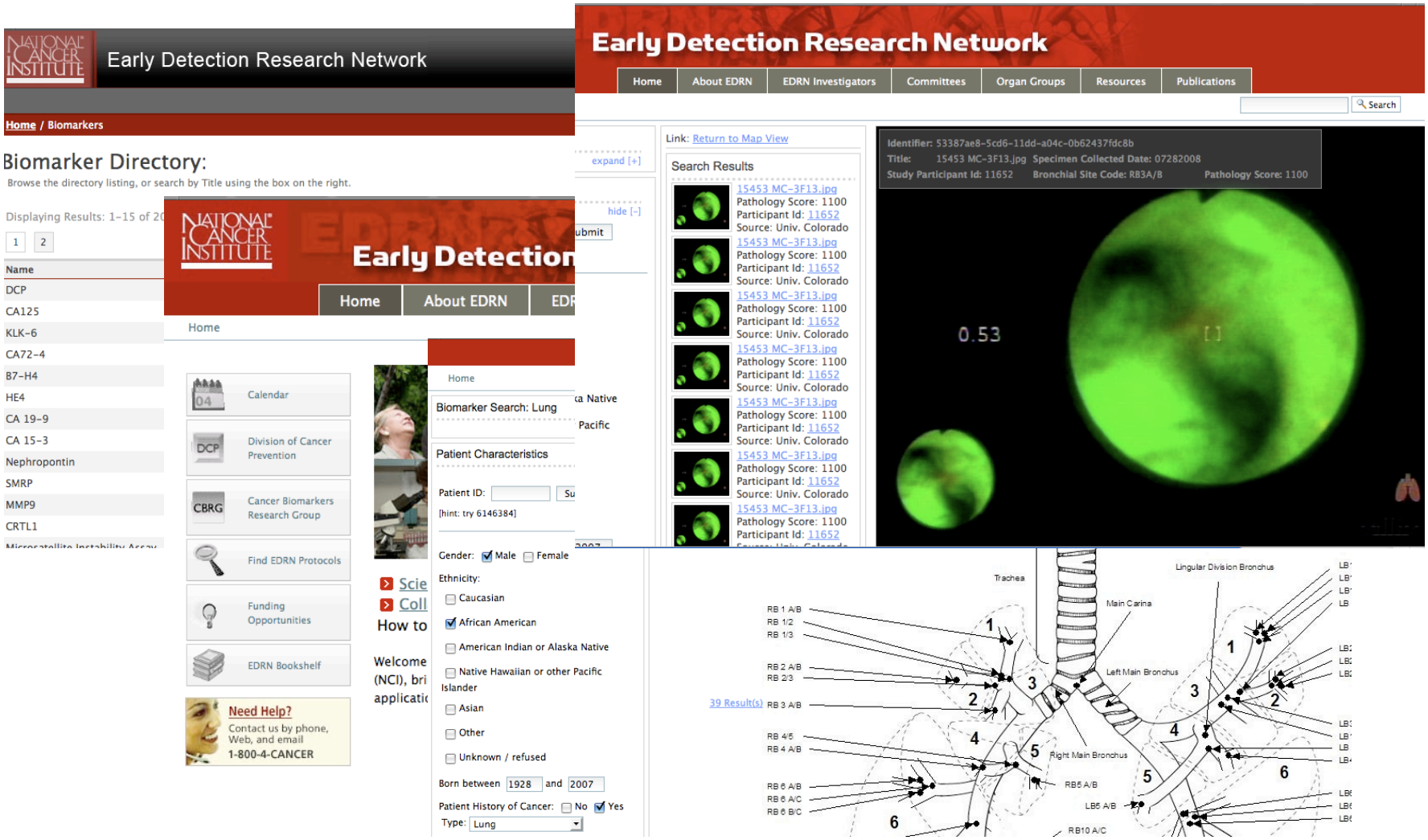
Release Policies

- Software
  - Documented at: https://cancer.jpl.nasa.gov/documents/applications/catalog-and-archive-service/deploying-the-ecas-to-operations
    - Describes how to deploy the eCAS software to operations
- Datasets
  - As datasets are acquired from PIs (via mobile eCAS, or via WebDAV, or via CD/DVD), there is a vetting process involving curation activities
    - Dataset-level Metadata curation
    - Metadata extraction for Product-level curation
    - Bulk Ingest, or product-by-product
    - Validation/Verification
      - Performed on tumor (dev) first, then performed on cancer (ops)
  - Datasets released currently along with software, but can be a separate activity as well

Data Sharing

- Defined in part by security policies and authorization at the eCAS/eCAS curation level
- Defined also in part through RDF dump of product/dataset metadata to portal and ability to download datasets
- Requests for eCAS data from private and research communities
  - Data Disclaimer
- eCAS meant to be a software/data enabler, not a lawyer

eCAS is just one of our operational systems: we are growing!



Get involved

- Check out our web presence:
  - http://cancer.jpl.nasa.gov/
- Check out our public portal (BETA):
  - http://edrn.jpl.nasa.gov
- Check out our facebook group:
  - http://www.facebook.com/group.php?gid=56938589930

Acknowledgements

- NCI
  - Don Johnsey
  - Christos Patriotis
  - Sudhir Srivastava
- Data Management and Coordinating Center (DMCC) at Fred Hutchinson Cancer Research Center
- EDRN IC
  - Dan Crichton, Andrew Hart, Steve Hughes, Heather Kincaid, Sean Kelly, John Tran, Thuy Tran

Questions